DOCUMENT RESUME

ED 237 550 TM 830 814

AUTHOR McLean, James E.; Lockwood, Robert E.
TITLE Sources of Error in an Angoff Type Standard Setting Process.
PUB DATE Nov 83
NOTE 13p.; Paper presented at the Annual Meeting of the Mid-South Educational Association (12th, Nashville, TN, November 16-18, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Standards; *Cutting Scores; Error of Measurement; Factor Analysis; Graduation Requirements; *Minimum Competency Testing; *Reliability; Secondary Education
IDENTIFIERS Alabama Basic Competency Tests; *Alabama High School Graduation Examination; *Angoff Methods; Standard Setting

ABSTRACT

The sources of variability in the Angoff standard-setting procedure, when applied to the Alabama High School Graduation Examination (AHSGE), were examined. The sources of variability examined are judges, rounds (replications), competencies (items), and interactions among these three sources. After training, the judges were given a statement of a competency, the item specifications for measuring that competency, and a number of sample items based on those specifications. The judges were then asked to estimate the number of items out of 100, similar to the sample items, that they expected a minimally competent student should answer correctly. The initial difficulty estimates (p-values) comprised round one, and the estimates after discussing these p-values comprised round two. The results from the ratings were analyzed using factor analyses of the three sources of variability. The analysis was repeated for the three subject areas of the AHSGE: reading, mathematics, and language. In general, the judge by competency interaction accounted for the largest portion of variance. It seems, therefore, that providing the judges with the actual difficulty levels of items has little impact on the standard setting process. (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED237550

SOURCES OF ERROR IN AN ANGOFF TYPE STANDARD SETTING PROCESS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

James E. McLean
The University of Alabama
P. O. Box 4006
University, Alabama 35486
205/348-7575

Robert E. Lockwood
Alabama State Department of Education
Room 607, State Office Building
Montgomery, Alabama 36130
205/832-3890

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. E. McLean

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper Presented at the Twelfth Annual Meeting
of the
Mid-South Educational Research Association
Nashville, Tennessee
November 16-18, 1983

TM 830 814

SOURCES OF ERROR IN AN ANGOFF TYPE STANDARD SETTING PROCESS

At least 35 states have implemented some type of minimum competency testing program. Many of these programs culminate in the establishment of minimum passing criteria for high school graduation. The recent resolution of Debra P. v. Turlington (1983) in Florida demonstrated that if certain standards are maintained in its development, a high school graduation test can overcome potential legal barriers.

Implementing high school graduation examinations requires establishing a minimum passing score or cutoff score. One of the more popular methods for establishing a minimum passing score is the procedure proposed by Angoff or some variation of that procedure. The Angoff procedure, first published as a foot- note in Angoff's contribution to Educational Measurement (Angoff, 1971), requires expert judges to determine what percentage of minimally qualified examinees are likely to answer an item correctly.

In this application, Angoff's procedure was modified in two ways. The first modification was having judges estimate what percentage of minimally qualified examinees should correctly answer all items measuring a competency. The second modification was incorporating item difficulty indices obtained during piloting as a moderating variable between the judges' first and second estimation process.

This paper examines the sources of variability in this procedure when applied to a high school graduation test. The

sources of variability examined are judges, rounds (replication), competencies (items), and interactions among these three sources.

Method

The three subject area examinations of the Alabama High School Graduation Examination (AHSGE) were subjected separately to an Angoff type process with three sets of 11 or 12 judges each. The judges were selected by individual superintendents based on criteria established by the researchers. The criteria for selection as a judge were: a) teaching the subject area they were to rate (reading, mathematics, or language), b) teaching at the high school level (9th - 12th grades), and c) teaching students who are marginally competent in these subject areas. These criteria were established to ensure that judges were both content area specialists and knowledgeable of borderline students in their subject area.

The judges received the following training in the standard setting process. First, a group leader provided the judges with a definition of a minimally competent student. Judges were then asked to describe a minimally competent student whom they had taught. The descriptions of the student were to include some behavioral characteristics. After this discussion, the judges were given a sample item and asked to estimate the number of similar items, out of 100, that a minimally competent student would be expected to answer correctly. The judges were then asked to discuss how they had arrived at their estimates. This process was repeated for several more sample items.

When the researchers were satisfied that the judges understood the rating process, the judges were given a statement of a competency, the item specifications for measuring that competency, and a number of sample items based on those specifications. The judges were then asked to estimate the number of items out of 100 similar to the sample items that they expected a minimally competent student should answer correctly.

After a brief check to ensure that the judges were recording their estimates correctly, the judges proceeded to provide estimates for each competency in their particular subject. The researchers observed the judges as they completed their ratings.

When all judges had completed their ratings, they were provided with difficulty estimates (p-values) for each of the competencies. These p-values were computed from the pilot test data by estimating the difficulty of each item in a random sample of students and averaging them for the competency. The judges were given an explanation of the p-values with the caution that they were computed for a random sample of all students in the State of Alabama and not just minimally competent students. A discussion followed about how these data might or might not affect the judges' estimates for each competency. The judges were then asked to reconsider their estimates and if they wanted to change an estimate to place the new estimate next to the original one. Their initial estimates comprised round 1 and their estimates after seeing the p-values comprised round 2.

The results from the ratings were analyzed using three-way factorial analyses of variance with judges (random), rounds (random), and competencies (random) as independent variables. The

analysis was repeated for reading, mathematics, and language. The primary reason for the analysis was to estimate the variance component for each potential source of variability. Estimates of the variance components are used to estimate the contribution of each source of variance to the total observed variance. The method used is one similar to that suggested by Cronbach, Gleser, Nanda, and Rajaratnam (1972).

Results

The results provided some clues to the magnitude of the contribution for the three sources of variability. In general, the judge by competency variance component accounted for the largest proportions of variance. The "round" component accounted for very little of the variance. It seems therefore, that providing the judges with the actual difficulty levels of items has little impact on the standard setting process. Estimates of the variance components for reading, mathematics, and language are shown in Tables 1, 2, and 3 respectively.

Discussion

A number of results are consistent across all three sets of judges for the three subject areas. Probably the most obvious of these is the small amount of change due to rounds. It should be noted that these small changes were noted even though the judges were given actual average p-values between the first and second rounds. Not only does rounds account for an almost nonexistent percentage of the total of the variance, the mean changes in the judges' average cutoff scores were small. As seen in Table 4, the

differences in average total test scores from round 1 to round 2 were .800, .013, and .739 for reading, mathematics, and language respectively. All of these differences are nonsignificant statistically at the .05 level.

In this particular application, omitting the second round would not have had an effect on the results of the standard setting process. At this time, we can not generalize beyond our data as many other factors may be involved (e.g., the differences between the judge's original estimates and the p-values). The results of this study do shed some light on the controversy concerning whether to provide judges with p-values or not.

A second finding of interest is that the variance components for the judges and the judges x competency interaction, when taken together, account for about two-thirds of the total variance. Still, the inter-judge reliabilities (coefficient alphas) computed using round 2 data are very high-- .93, .98, and .88 for reading, mathematics, and language respectively. Thus, it seems that although the judges do not necessarily agree on the difficulty of the items for each individual competency, they tend to agree on the overall levels of difficulties for each of the three subject area tests.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (pp. 508-600). Washington, D.C.: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: John Wiley & Sons, Inc.

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979, 1983).

TABLES

Table 1

Estimates of Variance Components for Reading

Source	SS	df	MS	Variance component	
				Estimate	Percentage
Judges	8688.94	10	868.89	26.60	33.7
Rounds	52.80	1	52.80	.22	.3
Competencies	6365.17	14	454.65	17.69	22.4
J x R	132.33	10	13.23	.56	.7
J x C	8724.70	140	62.32	28.74	36.4
R x C	113.29	14	8.09	.30	.4
J x R x C	678.58	140	4.85	4.85	6.1
Total	24755.81	329	-	78.96	100.0

Table 2

Estimates of Variance Components for Mathematics

Source	SS	df	MS	Variance component	
				Estimate	Percentage
Judges	36397.00	10	3639.70	61.75	40.0
Rounds	.03	1	.03	*	0.0
Competencies	22264.00	27	824.60	29.97	19.4
J x R	1063.70	10	106.37	3.27	2.1
J x C	24311.00	270	90.04	37.63	24.4
R x C	2427.40	27	89.91	6.83	4.4
J x R x C	3990.90	270	14.78	14.78	9.6
Total	90454.03	615	-	154.23	99.9

*Negative value, treated as zero (see Cronbach et al., 1972).

Table 3

Estimate of Variance Components for Language

Source	SS	df	MS	Variance component	
				Estimate	Percentage
Judges	2279.34	11	207.21	3.82	7.5
Rounds	49.14	1	49.14	*	0.0
Competencies	3674.33	14	262.45	7.12	13.9
J x R	414.63	11	37.69	1.95	3.8
J x C	9754.20	154	63.34	27.43	53.6
R x C	514.66	14	36.76	2.36	4.6
J x R x C	1307.08	154	84.88	8.49	16.6
Total	17993.38	359	-	51.17	100.0

*Negative value, treated as zero (see Cronbach et al., 1972).

Table 4

Estimates of Total Scores by Judges for Rounds 1 and 2

Subject	Mean		Mean difference
	Round 1	Round 2	
Reading	74.758	75.558	.800 NS
Mathematics	64.747	64.734	.013 NS
Language	79.839	80.578	.739 NS

NS--nonsignificant difference using given F-ratio.